Scalability of InfiniBand-Connected LNET Routers

Team Light Coral

Susan Coulter HPC-3
David Bonnie HPC-3
Christopher Hoffman HPC-3

Emily Baldwin Wheaton College
Matthew Schauer Georgia Institute of Technology
Jarrett Crews New Mexico Institute of Mining and Technology

Abstract

Lustre is one of the primary distributed file systems used at Los Alamos National Laboratory. It is generally connected using a high-speed InfiniBand interconnect in an arrangement known as a Lustre Network (LNET). Lustre configurations that assign distinct LNET routers for each Lustre file system are not cost-effective for large-scale cluster operation. To efficiently scale Lustre, it is in the interest of the high-performance computing (HPC) community to investigate the viability and performance of linking multiple Lustre file systems and their shared client nodes through a common set of LNET routers. In a test environment, we used the benchmarking tool IOR to measure the read and write bandwidth under varying conditions across a single LNET router connecting two Lustre file system servers and multiple client nodes. Modified conditions included the scale of access parallelism to the servers and the access block size. The benchmarks resulted in nearly uniform read and write speeds at a small scale of two Lustre servers. Performance scaled with the number of servers rather than the number of clients, indicating that the bottleneck was in the servers. Future research should investigate whether this trend persists in more complex systems with more servers and heavier traffic.

Methods

The test cluster is composed of 12 Lustre clients, 2 Lustre servers, and a single LNET router connected via FDR InfiniBand at 56 Gb/s. The router divides the cluster into two separately switched InfiniBand networks: one for the clients and one for the servers. All nodes contain two six-core hyperthreaded Intel Xeon processors and are running Linux kernel 2.6.32, patched for Lustre.

All nodes are diskless and booted over the network, so there is no persistence outside of the Lustre file systems themselves. Thus, both servers are set up identically, as are all the client nodes. All three types of node (server, router, and client) have the Lustre kernel modules installed and configured. On the servers and clients, the configuration tells each node which network it is on and the route to the other network. The configuration of the router tells it which network is attached to each of its interfaces. Beyond this base setup, the servers also need some extra utilities for creating and maintaining Lustre file systems. The clients are configured to automatically mount the file systems at startup.

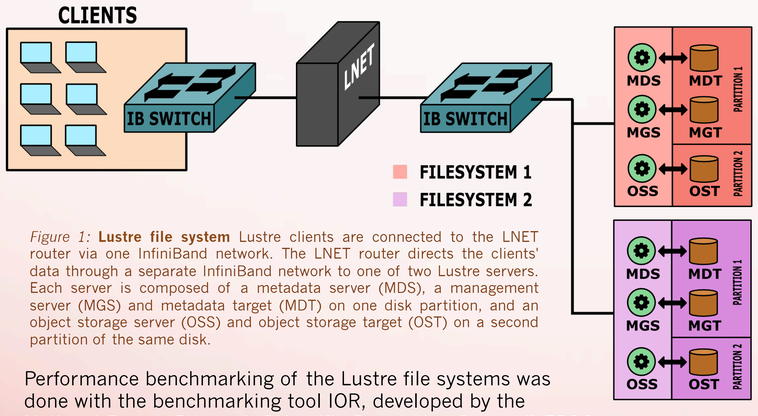

Figure 1: **Lustre file system** Lustre clients are connected to the LNET router via one InfiniBand network. The LNET router directs the clients' data through a separate InfiniBand network to one of two Lustre servers. Each server is composed of a metadata server (MDS), a management server (MGS) and metadata target (MDT) on one disk partition, and an object storage server (OSS) and object storage target (OST) on a second partition of the same disk.

Performance benchmarking of the Lustre file systems was done with the benchmarking tool IOR, developed by the National Energy Research Scientific Computing Center (NERSC). IOR writes or reads specified amounts of data to/from a mounted file system and reports bandwidth statistics for the transfer. IOR supports using MPI to distribute reads and writes across clients to simulate realistic scenarios of many clients reading and writing to the same server. To expedite testing, we created a script that would automate the sequence of a read, then a write, then a simultaneous read and write, using different combinations of clients, file sizes, and transfer block sizes. The combinations of parameters tested are summarized in Table 1. Each test was run eight times to reduce noise, and the mean and standard deviation of the test bandwidth were collected.

Benchmark Data

Number of Nodes	File Size/Process	Block Size	Processes/Node	Total Transfer Size
6	32 GB	1 GB	1	192 GB
6	32 GB	512 MB	1	192 GB
6	32 GB	2 KB	1	192 GB
6	1 GB	1 GB	24	144 GB
6	1 GB	512 MB	24	144 GB
6	1 GB	2 KB	24	144 GB

Table 1: **Varying test conditions** Combinations of variables for each test run measuring the read, write, and parallel read/write bandwidth and LNET router throughput for one and two Lustre file systems.

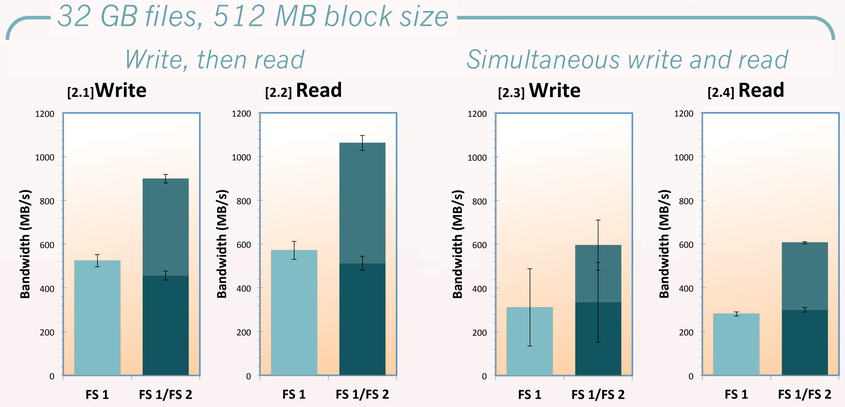

Figure 2: **Test using 32 GB files, transferring in 512 MB blocks** Sequential read then write operations are represented in the left two graphs, while the right two show simultaneous write/read. Standard deviation is denoted by the error bars for each file system.

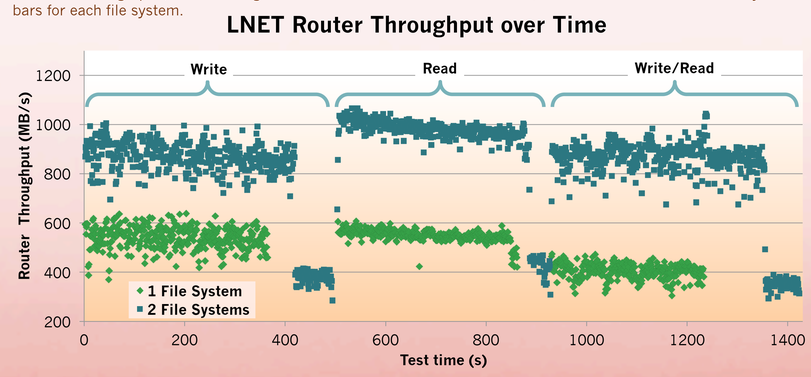

Figure 3: **LNET Router Throughput** This graph shows the router traffic for two POSIX operations on the Lustre file system. The low points of the two-file system data indicate when one file system finishes a job before the other.

Results and Analysis

The tests show that an overall speed of about 500 MB/s per five-disk server is obtainable. This result seems independent of the number of parallel operations and the size of individual transfers, although performance does degrade sharply for blocks smaller than 2 KB. Read and write speeds are comparable. Nearly double the speed of an individual operation is achieved when a read and write are done simultaneously.

Figure 2 shows the bandwidth difference between POSIX write and read operations on one and two file systems. The two file systems' aggregate bandwidth was equal to the sum of the individual file systems' bandwidths, as seen in Figures 2.1 and 2.2. This trend is also observed in tests with the router removed. When reading and writing simultaneously, shown in Figures 2.3 and 2.4, the overall bandwidth is reduced because the servers have to perform twice as many operations. The error bars on Figure 2.3, denoting standard deviation, are particularly large due to the read operation finishing before the write operation.

The LNET router throughput, in Figure 3, displays the three stages of bandwidth testing for both one and two file systems. The two file system operations are performed at twice the bandwidth of the single file system. This implies that two servers can operate as efficiently as a single server linked through one LNET router. At the conclusion of each operation in the two file system data, there is a significant drop in the router throughput; this occurs when one file system finishes its operation prior to the other.

Conclusions

The results demonstrate that the bottleneck in the system is in the servers, not the network, the clients, or the router. This confirms the original research proposition that multiple Lustre servers can be run through a single router. Furthermore, the load on the router was negligible during all of the benchmarks. The benchmark results indicate that there is no significant loss in bandwidth when writing to two file systems as opposed to one. This trend may not continue as the system is scaled, posing a potential problem to large clusters.

These results are strictly preliminary, considering the relatively small scale of the test environment with only two servers and six clients available to operate on each server. Nevertheless, this yields an initial understanding of the scalability of a single LNET router and future potential in more efficient Lustre file systems.

Future Work

The next step for this research is to test the scalability of LNET routers to several or tens of servers. It is conceivable that entire clusters could be served with a single router. It would also be interesting to see the effects of more complex setups, such as:

- Lustre file system components (MDT, OST, etc.) on different servers
- heterogeneous networks connected partially with InfiniBand and partially with Ethernet or some other interconnect
- multiple Lustre networks with varying numbers of servers on each
- multiple routers connecting many Lustre networks

These setups would require much more equipment and setup than our situation permits, but they are necessary to confirm the robustness of Lustre routing.

Emily Baldwin
baldwinemb@gmail.com
Jarrett Crews
jarrett.crews@gmail.com
Matthew Schauer
matthew.schauer.x@gmail.com

LA-UR-14-26013